Description 4/PRTS

Method for encoding an XML-based document

The invention relates to a method for encoding an XML-based document (DOC) which includes contents corresponding to an XML schema language definition. It further relates to a corresponding decoding method as well as to corresponding encoding and decoding devices.

XML (= Extensible Markup Language) is a language which enables a structured description of the contents of a document to be produced by means of XML schema language definitions. A more detailed description of the XML schema and also of the structures, data types and content models used therein can be found in the references [1], [2] and [3].

Methods for encoding XML-based documents in which the document is converted into a coded binary representation are known from the prior art. Methods for encoding and decoding XML-based documents are described for example in document [4] which was produced in the course of the development of an MPEG-7 coding standard.

The methods for generating a binary representation of XML-based documents that are known from the prior art have disadvantages when it comes to the encoding of "complex type" data types with the "mixed" content model, because in addition to elements said data types can include textual contents which can, however, only be reconstructed by the decoding of the entire data stream. A more detailed description of the "complex type" data type and of the "mixed" content model can be found in document [1].

The object of the invention is therefore to create a method for encoding XML-based documents which enables easier access to coded textual contents of the "complex type" data type with the "mixed" content model.

This object is achieved by the independent claims. Developments of the invention are defined in the dependent claims.

With the encoding method according to the invention, a coded binary representation of an XML-based document is created in that the contents of the document are assigned binary structure codes by way of coding tables, with structure codes being assigned to textual contents of a "complex type" data type with the "mixed" content model. The structure codes are the schema branch codes (SBC) defined in section 7.6.1 of document [4]. As a result of the assignment of structure codes to contents of the document as described in [4], the location of said contents in the structure of the XML documents can be signaled or addressed.

The invention is essentially characterized in that the textual content of a "complex type" data type with "mixed" content model is treated as an element declaration in the type definition during the code assignment. Accordingly, for the purposes of encoding, as well as the declared elements a specified structure code is also assigned in addition to the textual content in a type definition when a mixed content model is defined for the type. By this means textual contents in the coded data stream are addressed, with the result that said contents can be accessed without the need to decode the entire data stream.

In a preferred embodiment of the encoding method according to the invention, the structure codes are assigned to the textual

contents of a "complex type" data type with "mixed" content model exclusively via OperandTBC coding tables. Said coding tables specify the codes of what are referred to as the OperandTBCs, i.e. the so-called TBCs (TBC = Tree Branch Code) of the so-called operand nodes. A detailed description and definitions of the OperandTBCs and operand nodes can be found in sections 7.6.1 and 7.6.5.2 of document [4].

In a particularly preferred embodiment, "position codes" are also assigned to the textual contents of a "complex type" data type with the "mixed" content model. These codes are the position codes described in more detail in section 7.6.5.5 of document [4]. Since a plurality of textual contents may be contained in a "complex type" data type with the "mixed" content model, said position codes serve to transmit the information indicating at which position the textual contents are located within the data type.

In a particularly preferred embodiment, "single element position codes" and/or "multiple element position codes" are used for the assignment of the "position codes". Said position codes are described in more detail in publication [4], section 7.6.5.5. Single element position codes are used in particular when no "model group" can occur more often than once in the type definition of the "complex type" in the XML schema definition. A definition of the "model group" can be found in document [2]. In this case the single element position code determines the position of a content in relation to a particular particle in an instantiation of a data type. A definition of particles can also be found in document [2]. The single element position code is encoded on the assumption that the textual content is declared a maximum of MPA+1 times, where MPA denotes the number of all particle instantiations possible in this data type. A multiple element position code is used

when "model groups" can occur more often than once in the definition of the "complex type" in the XML schema definition. The multiple element position code is encoded on the assumption that a total of 2*MPA+1 positions can be addressed, with this code reflecting the position of the content in relation to all particles in an instantiation of a data type.

In a further preferred embodiment, the position codes are encoded using codes of variable length, in particular using the code vluimsbf5, which is described in document [4], section 4.3.

In addition to the above described encoding method the invention further comprises a decoding method by means of which a binary representation of an XML-based document encoded according to the above described encoding method is decoded. With said decoding method, binary representations of textual contents of a "complex type" data type with the "mixed" content model which were assigned structure codes (SBC) during the encoding are converted into the textual contents of the XML-based document which were assigned to the structure codes (SBC).

Analogously to the encoding method, in a preferred embodiment the assignment is effected by means of structure codes (SBC) by way of OperandTBC coding tables.

In a preferred embodiment, binary representations of textual contents of a "complex type" data type with the "mixed" content model, addressed by means of "position codes", are also converted into textual contents at the assigned position. In this case the "position codes" can in turn comprise "single element position codes" and/or "multiple element position codes". These position codes are the same position codes as

defined in relation to the encoding method. Analogously to the encoding method, the "position codes" can also be encoded using codes of variable length, said codes being decoded during the conversion of the position codes into textual contents. The position codes are preferably encoded using the code vluimsbf5.

In addition to the above described encoding and/or decoding methods, the invention further comprises an encoding and decoding method which comprises the encoding method according to the invention and the decoding method according to the invention.

The invention further relates to a device for encoding XML-based documents by means of which the encoding method according to the invention can be performed, said device comprising a storage means in which at least one assignment of a textual content of a "complex type" data type with the "mixed" content model to a structure code is stored. In an analogous manner the invention relates to a device for decoding a coded binary representation of an XML-based document, said device being configured in such a way that the decoding method according to the invention can be performed. The device comprises a storage means in which at least one assignment of a structure code to a textual content of a "complex type" data type with the "mixed" content model is stored.

The invention further relates to a device for encoding and decoding an XML-based document, comprising the above described encoding device according to the invention and the above described decoding device according to the invention.

Exemplary embodiments of the invention are explained below with reference to the attached drawings, in which:

Figure 1 shows a schematic diagram of an encoding and decoding system according to the invention having an encoder and a decoder;

Figure 2 shows an XML schema definition in which, among other things, a "complex type" data type with "mixed" content model is defined;

Figure 3 shows an XML document in which an element "MixedElement" declared in the XML schema definition of Fig. 2 is instantiated;

Figure 4 shows a graphical representation of the structure of the element "MixedElement" instantiated in the XML document of Fig. 3;

Figure 5 shows a table serving to explain the assignment of structure codes for "complex type" data types with "mixed" content model; and

Figure 6 shows a table serving to explain the assignment of "position codes" for "complex type" data types with "mixed" content model.

Figure 1 shows by way of example an encoding and decoding system comprising an encoder ENC and a decoder DEC by means of which XML documents DOC are encoded and decoded respectively. Both the encoder and the decoder have what is referred to as an XML schema S in which the elements and types of the XML document which are used for communication are declared and defined. Code tables CT are generated in the encoder and decoder from the schema S via corresponding schema compilations SC. When the XML document DOC is encoded, binary codes are assigned to the contents of the XML document via the code

tables. By this means a binary representation BDOC of the document DOC is generated, which binary representation can be decoded again in the decoder with the aid of the code table CT.

The method according to the invention is characterized in that textual contents of a "complex type" data type with the "mixed" content model are assigned binary structure codes. This enables the textual data to be filtered out from the binary representation BDOC without the need to decode the entire binary representation BDOC.

Figure 2 shows by way of example a schema S, an element with the name "Example" being declared in this schema in lines 4 to 10, said element in turn containing an element with the name "MixedElement" of the type "MixedType". The type "MixedType" is defined in lines 12 to 17. Said type is a "complex type" data type with the content model "mixed", as can be derived in particular from line 12. The type "MixedType" contains two elements with the name "firstElement" and "secondElement", both of which are of the "string" type.

Figure 3 shows an instantiation of the element "MixedElement" in an XML document. Since the "mixed" content model can include textual contents in the form of strings, textual contents can occur before, after or between the first and second elements "firstElement" and "secondElement". A total of three textual contents occur in the example shown in Figure 3.

The structure of the element "MixedElement" which is instantiated in Figure 3 is shown again in Figure 4, this time in graphical form as a tree structure. From the topmost MixedElement/MixedType node there branch off at a first hierarchy level five further nodes which contain both the textual contents and also the elements "firstElement" or, as

the case may be, "secondElement". At a second hierarchy level the elements "firstElement" and "secondElement" further contain the corresponding contents "Content of firstElement" and "Content of secondElement" respectively.

Any document based on the XML language can be represented by what is referred to as a tree structure, the contents of the XML document forming nodes in the tree structure and what are referred to as "context paths" leading to said nodes. Binary structure codes are assigned to the nodes of the tree structure during the encoding.

According to the prior art, for the element node "MixedElement" shown in Figure 4 a structure code is assigned in each case to the parent node and also to the elements "firstElement" and "secondElement". In this case the parent node is the node which is connected to the node of the element "MixedElement" in the next-higher hierarchy level. In contradistinction hereto, with the method according to the invention a structure code is not only assigned to the parent node and to the elements "firstElement" and "secondElement"; rather, a structure code is also assigned to the textual content. This is illustrated in Figure 5, where the code 00 is assigned to the parent node, the code 01 to the textual content and the codes 10 and 11 to the "firstElement" and the "secondElement" respectively.

In the method according to the invention it is also possible to assign "position codes" to the individual textual contents, as shown in Figure 6. Since textual contents can occur at a total of three positions, three "position codes" are required for this purpose, the codes 00, 01 and 10 being used according to Figure 6.

List of references:

[1] http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/

[2] http://www.w3.org/TR/2001/REC-xmlschema-1-20010502/

[3] http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/

[4] ISO/IEC FDIS 15938-1 "Information Technology - Multimedia Content Description Interface - Part 1: Systems", Geneva 2002